A Computable Theory for Learning Bayesian Networks based on MAP-MDL Principles

Heping Pan
International Institute for Financial Prediction
School of Information Technology & Mathematical Sciences
University of Ballarat
Mt Helen, VIC 3353, Australia
www.iifp.net, h.pan@iifp.net or h.pan@ballarat.edu.au

Daniel McMichael
The Business Intelligence Group
CSIRO Mathematical and Information Sciences
Cornish Building, Gate 5, Waite Rd
Urrbrae, SA 5064, Australia
http://www.cmis.csiro.au/bi, Daniel.McMichael@csiro.au

***Abstract*— Bayesian networks provide a powerful intelligent information fusion architecture for modeling probabilistic and causal patterns involving multiple random variables. This paper advances a computable theory of learning discrete Bayesian networks from data. The theory is based on the MAP-MDL principles for maximizing the joint probability or interchangeably miniziming the joint description length of the data and the Bayesian network model including the network structure and the probability distribution parameters. The computable formalisms for the data likelihood given a Bayesian network structure, the description length of a structure, and the estimation of the parameters given a structure are derived. EM algorithms are constructed for handling incomplete and soft data.**

I. Introduction

By exploiting the factorization of the joint probability distribution, Bayesian networks provide a powerful architecture for information fusion of multiple disparate variables. Learning Bayesian networks has been a central area in Bayesian network research in recent years because it serves two important purposes: to overcome the bottleneck of constructing Bayesian networks combining expert knowledge and statistical data, and to enable discovery of causalities for rational, reliable and understandable modeling. The problem can be divided into two subproblems: structural learning and parametric learning. In general, structural learning is a NP-hard problem since the space of all possible structures has a size exponential to the number of variables and the number of discrete states of each variable [1]. The directional acyclicity is also a subtle constraint. Incomplete or soft data renders closed-form solutions nonexistent.

A Bayesian network BN, also called causal probabilistic network, for a problem domain under consideration is a *directed acyclic graph (DAG)* G representing the full joint probability distribution $P(\mathbf{V})$ over the set of variables $\mathbf{V}$ for the problem domain

$$BN = (G,\, \mathbf{P}) = (\mathbf{V}, \mathbf{L}, \mathbf{P}) \tag{1}$$

where $\mathbf{V} = \{V_1, V_2, \ldots, V_n\}$, G is a DAG which is defined by putting each variable V of $\mathbf{V}$ as a node, and each causal relation from each of V's parents to V as a directed link. We therefore shall use the term variables and nodes interchangeably. $\mathbf{L}$ denotes the set of all the directed links while $\mathbf{P}$ is a set of conditional probability distributions associated with every node $V \in \mathbf{V}$ given V's parents Γ_V^+. We use lower-case letters such as v, v_j, to denote *instantiation* of the corresponding variables to an actual *value* or *state*. Written in mathematical forms, we have

$$G = (\mathbf{V}, \mathbf{L}) \tag{2}$$

$$\mathbf{L} = \{(U, V) \text{ for } V \in \mathbf{V}, U \in \Gamma_V^+\} \subseteq \mathbf{V} \times \mathbf{V} \tag{3}$$

$$\mathbf{P} = \{P(V|\Gamma_V^+) \text{ for } V \in \mathbf{V}\} \tag{4}$$

Using Bayes chain rule, it can be shown that if no directed cycles exist among all the causal relations between variables of $\mathbf{V}$, the full joint $P(\mathbf{v})$ is equivalent to the set $\mathbf{P}$ defined by (4), and

$$p(\mathbf{v}) = \prod_{V \in \mathbf{V}} p(v|\gamma_V^+) \tag{5}$$

where γ_V^+ denotes an instantiation of Γ_V^+. This factorization is fundamental and all existing inference methods were actually derived by manipulations of this equation.

Learning a Bayesian network from data in general means to determine from a given data set the structure of the network, i.e. the set of causal links among variables, and the probability parameters associated with the

AI'2005 Workshop on Learning Algorithms for Pattern Recognition, UTS, pp. 24-41.

structure. Although ideally the structure and parameters should be learned simultaneously, finding the optimal structure of the network is the most difficult part of the whole problem. First of all, the possible number of alternative structures is exponential to the number of parents for each variable. This renders the structural learning a NP-hard problem.

The problem of learning Bayesian networks can now be defined as follows: A data set $\mathbf{D}$ is given, which is a collection of data cases $\mathbf{D} = \{D_k | k = 1, 2, \ldots, N\}$ where each D_k is a data case. In general, if we do not want to specify which data case, we may use $D \in \mathbf{D}$ to denote a single data case. A model space $\mathcal{M}$ exists, which is a set of all possible Bayesian networks that can be hypothesized for describing the given data set $\mathbf{D}$: $\mathcal{M} = \{M\}$ where each model M is a complete Bayesian network

$$M = (\mathbf{V}, \mathbf{L}, \mathbf{P}) = (S, \Theta) \tag{6}$$

In general, we assume the variable set $\mathbf{V}$ is predefined. S denotes the structure of the network referring to the set of directed links $\mathbf{L}$, and Θ denotes the vector of all the parameters specifying all the conditional probability tables $\mathbf{P}$ given the structure S.

In the literature, [2] provides an in-depth tutorial on Bayesian approach to learning Bayesian networks; [3] offers a comprehensive guide to the literature on learning probabilistic networks from data; [4] is a recent review paper on learning Bayesian networks; [5] collects introductory surveys and papers describing recent advances; [6] extends to hybrid Bayesian networks. *Structure learning* is a term used for learning the structure of Bayesian networks from data. The difficulties in structure learning arise from (1) the structure space - the set of all possible different structures for a target Bayesian network to be learned from a given data set - is not continuous; (2) the size of the structure space is exponential relative to the number of variables; (3) for each possible network structure, the acyclicity of directed links has to be guaranteed; (4) equivalence classes of structures in terms of conditional independence properties have to be considered in enumerating possible different structures or in attempting to learn a subset of reasonable structures rather than a single best one.

The Bayesian approach of probability and statistics includes Bayesian averaging and maximum aposterior probability (MAP) principle. The MAP approach to learning the structure of multiply-connected networks was originally proposed by Cooper and Herskovits [7]. Their approach tries to find the most probable network using the Bayesian score, which is a product of the likelihood function of the data given a network structure and the prior probability of the structure. Like all Bayesian approaches, they must assume a prior distribution over the structure space. However, they took this prior to be uniform, which rendered the approach closely equivalent to ML estimation. In other words, by choosing the uniform prior, their approach would prefer a more accurate network irrespective of the structure complexity. The Bayesian approach to structural learning was worked out in different forms by many other researchers [4]. The general case for structure learning with the exponential family is worked through by Geiger and Heckerman [8].

A natural way to avoid explicitly defining the structure prior is the use of the minimum description length (MDL) principle [9]. With the MDL score, the prior of a hypothesized network structure is replaced by the description length of the structure. The most important point here is that this length is computable. Methods for structure learning using the MDL score were developed by Suzuki [10], Lam and Bacchus [11] and Bouckaert [12].

Friedman and Koller [13] provided an efficient algorithm for Bayesian model averaging in discrete Bayesian networks for a given variable ordering and provided an MCMC process for averaging over model orderings. Their approach provides a smoother search and is much faster than Madigan and York's MC^3 algorithm [14]. Green and others have investigated structure estimation in both continuous and discrete graphical models and have provided a reversible jump algorithm that enables the marginals of any structural or model parameter to be evaluated [15], [16]. These procedures still scale poorly, except in special cases, where for example, Taskar *et al.* have provided efficient algorithms for associative markov networks [17].

Structure learning in the presence of incomplete data and hidden variables have been attempted. In particular, Friedman [18] has attempted to extend the EM algorithm for parameter learning to structure learning in a method he called *Structural EM*. Roughly speaking, Structural EM performs searches in the joint space of (Structure $\times$ Parameters). At each step, it can either find better parameters for the current structure, or select a new structure. The former case is a standard "parametric" EM step, while the later is a "structural" EM step. However, since this joint space is not continuous, EM is not likely to perform as well as desired. The discontinuity of the joint space is due to the discontinuity of the structure space

which renders the "structural" EM step less significant of the original EM spirit than the "parametric" EM step.

Genetic algorithms have been applied to structure search in Bayesian networks, both for variable ordering [19], [20] and for direct structure estimation [21], [22].

On the basis of Bayesian networks, neural networks and multivariate time series analysis, Pan [23] proposed a new type of nonlinear dynamic Bayesian networks - Super Bayesian Influence Networks (SBIN) for modeling and predicting stochastic and chaotic patterns in multivariate time series. Essentially a SBIN is a dynamic Bayesian network whose nodes correspond to a set of time series. The nonlinear probability distribution for a variable (node) given its parents is provided by a Probability Ensemble of Neural Networks (PENN). Note that variables in a SBIN are generally continuous. A PENN can represent any probability distribution of continuous variables. However, learning the structure of a SBIN is equal to detecting the conditional influences among the multivariate time series.

From the next section on, we shall present a computational theory for learning Bayesian networks from data. The theory is based on an integrated criterion of maximizing the joint probability or interchangeably miniziming the joint description length of the data and the Bayesian network model including the network structure and the probability distribution parameters.

II. A Joint Criterion Based On MAP and MDL

The Bayesian approach is a well-founded and practical method for selecting the best among alternative models given a data set. The basic idea is to select the model which maximizes the aposterior probability of the model given the data, i.e.

$$M_{best} = \arg\max_{M} P(M|\mathbf{D}) \tag{7}$$

which is equivalent to

$$M_{best} = \arg\max_{M} P(\mathbf{D}, M) \tag{8}$$

Therefore, we shall call $P(\mathbf{D}, M)$ the MAP score - the objective function - of the model M given the data set $\mathbf{D}$. Using Bayes' rule and the model formation (6), the MAP score can be elaborated as

$$\begin{aligned} P(\mathbf{D}, M) &= P(\mathbf{D}|M)P(M) \\ &= P(\mathbf{D}|\Theta, S)P(\Theta|S)P(S) \end{aligned} \tag{9}$$

In pure structure learning, we may only want to compare alternative structures without estimating the parameters Θ. Notice that given a structure S, the parameters Θ is then defined in a parametric space Θ_S depending on S, i.e. $\Theta \in \Theta_S$. Therefore, for structural learning only, we may define the Bayesian score for a given structure S as

$$\begin{aligned} P(\mathbf{D}, S) &= \int_{\Theta_S} P(\mathbf{D}, \Theta, S) d\Theta \\ &= \int_{\Theta_S} P(\mathbf{D}|\Theta, S)P(\Theta|S)P(S) d\Theta \\ &= \left[\int_{\Theta_S} P(\mathbf{D}|\Theta, S)P(\Theta|S) d\Theta\right] P(S) \\ &= P(\mathbf{D}|S)P(S) \end{aligned} \tag{10}$$

where $P(\mathbf{D}|S)$ is the likelihood of the data set $\mathbf{D}$ given the structure S

$$P(\mathbf{D}|S) = \int_{\Theta_S} P(\mathbf{D}|\Theta, S)P(\Theta|S) d\Theta \tag{11}$$

A. An Information-Theoretical MDL Criterion

Learning a Bayesian network from a data set can be regarded as a problem of explaining the given set of statistical data using a learned Bayesian network as a model. By explanation, we naturally presume that there should be some structure underlying the data, and some redundancies complementing the structure. A thorough understanding of the data set would mean that we may give a description of the structure and redundancies which together, in turn, could determine the data set completely. Our purpose in general is only to reach a perfect, nonredundant description of the data by removing all redundancy. In this sense, there is a unique criterion for model selection and estimation, namely to consider the total number of bits - binary digits - with which the data set and the model can be written down completely. This number is called the total description length of the data including its explaining model. The shorter this description length is, the better the model is. The best model of all alternative models will be the one with the shortest total description length. This is the intuitive formulation of Minimum Description Length (MDL) principle [24], [9]. Pan and Förstner [25] applied this principle for automatic architecturing of feedforward neural networks, a problem closely resembles the Bayesian network learning. Learning Bayesian networks from data using MDL score has been investigated by Suzuki [10], Lam and Bacchus [11] and Bouckaert [12]. Methods developed by these authors differ mainly on the algorithmic level, namely on the strategies and techniques used for global minimization of the MDL score.

There is a general assumption underlying the application of MDL principle, namely there is a description language, denoted by $\mathcal{L}$, of the given problem domain. We shall use $\mathcal{L}(X)$ to denote the complete description of X in the description language $\mathcal{L}$, and $L(X)$ the length (total number of bits) of $\mathcal{L}(X)$. With these concepts, we require the description language $\mathcal{L}$ to have the following properties: completeness, efficiency, computability, stability.

Given a data set $\mathbf{D}$ out of the data space $\mathcal{D}$, the MDL principle selects the best model M_{best} out of the model space $\mathcal{M}$ with

$$M_{best} = \arg \min_{M \in \mathcal{M}} L(D, M) \tag{12}$$

In Bayesian network learning, the description language $\mathcal{L}$ is made up of Bayesian probability theory and directed acyclic graph theory. It can be shown that this language possesses the property of completeness, efficiency and stability to a sufficient extent, and that of computability partially. Computability depends very much on the assumptions we use for modeling probability distributions.

The model space $\mathcal{M}$ is a Cartesian product of the structure space $\mathcal{S}$ and the parameter space Θ_S given structure S

$$\mathcal{M} = (\mathcal{S}, \Theta_S) = \mathcal{S} \times \Theta_S, \qquad S \in \mathcal{S} \tag{13}$$

The joint description length of a given data set $\mathbf{D}$ and a model - a Bayesian network M - can be defined as

$$\begin{aligned} L(\mathbf{D}, M) &= L(\mathbf{D}, \Theta, S) \\ &= L(\mathbf{D}, \Theta|S) + L(S) \\ &= L(\mathbf{D}|\Theta, S) + L(\Theta|S) + L(S) \\ &= L(\mathbf{D}|\Theta) + L(\Theta|S) + L(S) \end{aligned} \tag{14}$$

For structural learning purpose, we may only need to evaluate

$$L(\mathbf{D}, S) = L(\mathbf{D}|S) + L(S) \tag{15}$$

This is in accordance with $P(\mathbf{D}, S)$ in (10).

B. An Integrated Criterion Mixing MAP and MDL Scores

According to the information theory, we can then establish an integrated criterion which selects the optimal structure S_o by

$$\begin{aligned} S_o &= \arg \max_{S \in \mathcal{S}} [P(\mathbf{D}|S)P(S)] \\ &= \arg \min_{S \in \mathcal{S}} [L(\mathbf{D}|S) + L(S)] \end{aligned} \tag{16}$$

where the corresponding terms of probability and description lengths are convertible via the following equations

$$L(\mathbf{D}|S) = -\log_2 P(\mathbf{D}|S) \tag{17}$$

$$L(S) = -\log_2 P(S) \tag{18}$$

This integrated criterion has two equivalence forms:

The first form is an elaboration of MAP which maximizes the joint probability

$$P(\mathbf{D}, S) = P(\mathbf{D}|S)P(S) = P(\mathbf{D}|S)2^{-L(S)} \tag{19}$$

where $L(S)$ is used to compute $P(S)$ by inversing equation (18) and

$$P(S) = 2^{-L(S)} \tag{20}$$

because there is no generally valid assumption about the original prior probability $P(S)$ of a particular structure S. But on the other hand, whenever the structure S is hypothesized, the likelihood $P(\mathbf{D}|S)$ of the data set $\mathbf{D}$ given the structure S may well be computable as shown by [7].

The second form is an elaboration of MDL which minimizes the joint description length

$$\begin{aligned} L(\mathbf{D}, S) &= L(\mathbf{D}|S) + L(S) \\ &= -\log_2 P(\mathbf{D}|S) + L(S) \end{aligned} \tag{21}$$

where $P(\mathbf{D}|S)$ is used to compute $L(\mathbf{D}|S)$ by equation (17) because the description $\mathcal{L}(\mathbf{D}|S)$ corresponds to the residuals after fitting the structure S to the data $\mathbf{D}$, and in general the probability $P(\mathbf{D}|S)$ as the likelihood of $\mathbf{D}$ given S is computable.

III. Computing the Description Length of a Network Structure

In reality, since there is apparently no plausible way of defining the prior $P(S)$ directly, we shall only consider the description length $L(S)$ which is computable because the structure S being an acyclic directed graph of discrete variables is a finite discrete data structure.

We still assume there are totally n variables that are given in an original fixed order, so each variable V_i can be referred by an integer index i. To encode a particular Bayesian network structure S, the following information is necessary and sufficient: the list of discrete states for each variable, whose description length is denoted by $L(\Omega_i)$, a list of parents for each variable, whose description length is denoted by $L(\Gamma_i^+)$, a conditional

probability tables for each variable given its parents, whose description length is denoted by $L(P(V_i|\Gamma_i^+))$.

Therefore the total description length of a Bayesian network structure S is

$$\begin{aligned} L(S) &= \sum_{i=1}^{n} \Big[L(\Omega_i) + L(\Gamma_i^+) + L(P(V_i|\Gamma_i^+)) \Big] \\ &= L_\Omega + L_\Gamma + L_P \end{aligned} \quad (22)$$

where

$$L_\Omega = \sum_{i=1}^{n} L(\Omega_i) \quad (23)$$

$$L_\Gamma = \sum_{i=1}^{n} L(\Gamma_i) \quad (24)$$

$$L_P = \sum_{i=1}^{n} L(P(V_i|\Gamma_i^+)) \quad (25)$$

The simple expressions for L_Ω, L_Γ and L_P are as follows

$$L_\Omega = \sum_{i=1}^{n} a(r_i - 1) \quad (26)$$

$$L_\Gamma = \sum_{i=1}^{n} b q_I \quad (27)$$

$$L_P = \sum_{i=1}^{n} c(r_i - 1) \prod_{V_j \in \Gamma_i^+} r_j \quad (28)$$

where a, b, c are constants: the number of bits required to encode, respectively, an integer index in the range of 1 to r_i for a state list, an integer index in the range of 1 to n for a parent list, and a probability as a real number whose precision is dependent on the application requirement. r_i and q_i are introduced previously, meaning the number of states for variable V_i and the number of parents of V_i.

Equation (22) is general when the structure learning also includes the variable discovery, so that the variable set $\mathbf{V}$ may vary with different alternative structures. $L(\Omega_i)$ can be dropped normally because we assume the set of variables $\mathbf{V}$ is fixed for all different alternative structures. Therefore, equation (22) may be rewritten as

$$L(S) = L_\Gamma + L_P \quad (29)$$

The integrated criterion seeks to minimize the description length $L(S)$, which tends to favor networks in which there are sparser connections among variables, or in other words, the nodes have a smaller number of parents (referring to L_Γ), and also less connections between nodes taking on a large number of states (referring to L_P).

IV. Computing Data Likelihood and Parameter Expectation

We now consider how to compute the likelihood $P(\mathbf{D}|S)$ of the data set $\mathbf{D}$ and and the expectation of parameters $E[\theta_{ijk}|\mathbf{D}, S]$ given a hypothesized structure S. For a discrete Bayesian network, a general model for multivariate joint distribution is the *multinomial distribution*, which is used here for computing $P(\mathbf{D}|S)$ and $E[\theta_{ijk}|\mathbf{D}, S]$. Cooper and Herskovits (1992) [7] derived $P(\mathbf{D}, S)$ and $E[\theta_{ijk}|\mathbf{D}, S]$ for the complete hard data type under multinomial distribution. We shall first follow their derivation for the data type 1 (complete and hard data set), and then extend the results to more general data types including incomplete and soft data sets.

A. Estimating Data Likelihood $P(\mathbf{D}|S)$ with Complete Hard Data

For the predefined variable set $\mathbf{V}$, a hypothesized structure S, and a given data set $\mathbf{D}$, we assume, for the time being, following [7], that A1: The variables in $\mathbf{V}$ are discrete (we only consider discrete Bayesian networks); A2: Cases in $\mathbf{D}$ occur independently, given $\mathbf{V}$, S, and the parameter space Θ_S; A3: The data set $\mathbf{D}$ is of the data type 1, i.e. contains only complete and hard data cases; A4: Before observing $\mathbf{D}$, we are indifferent to the prior probability $P(\Theta|S)$ of the parameters Θ given the structure S.

Before we elaborate, we need more notation. Let V_i denotes the i-th variable of $\mathbf{V}$, and V_i has r_i states. v_{ik} denotes the k-th state of the V_i. Let $\mathbf{D}$ be a data set of m cases, D_l denotes the l-th case of $\mathbf{D}$. Variable V_i in the structure S has a set of parents Γ_i^+. Let γ_{ij}^+ denote the j-th unique instantiation of Γ_i^+, which is a configuration of the parent variables of V_i, and there are q_i such instantiations, i.e. $q_i = |\Gamma_i^+|$. Now define N_{ijk} to be the number of cases in $\mathbf{D}$ in which variable V_i is instantiated to the state v_{ik} and its parent set Γ_i^+ is instantiated to the configuration γ_{ij}^+, i.e.

$$N_{ijk} = |\{D \in \mathbf{D} | V_i \in D = v_{ik}, \Gamma_i^+ \in D = \gamma_{ij}^+\}| \quad (30)$$

and also let $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ Given the structure S, the form of the conditional probability parameters Θ is defined. Let θ_{ijk} denote the conditional probability

$$\theta_{ijk} = P(V_i = v_{ik} | \Gamma_i^+ = \gamma_{ij}^+, \Theta) \quad (31)$$

and let $\theta_{ij} = (\theta_{ij1}, \ldots, \theta_{ijr_i})$ which represents a probability distribution of V_i given a configuration of its

parents. Evidently, we have

$$0 \leq \theta_{ijk} \leq 1 \tag{32}$$

$$\sum_{k=1}^{r_i} \theta_{ijk} = 1 \tag{33}$$

Let $\theta_i = (\theta_{i1}, \ldots, \theta_{iq_i})$ which represents the full conditional probability distribution (table) of V_i. With this notation, for $|\mathbf{V}| = n$, we have an explicit form of Θ given S as

$$\begin{aligned} \Theta &= (\theta_1, \theta_2, \ldots, \theta_n \quad) \\ &= (\theta_{11}, \ldots, \theta_{1q_1}, \theta_{21}, \ldots, \theta_{2q_2}, \ldots, \theta_{n1}, \ldots, \theta_{nq_n}) \end{aligned} \tag{34}$$

Following [7] we arrive at a closed-form solution for $P(\mathbf{D}|S)$:

$$P(\mathbf{D}|S) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{35}$$

Incorporating Prior Information about the Parameters Θ

An ordinary form of the prior information about the parameters Θ given the structure S is the set of counts

$$\{N'_{ijk} | i = 1, \ldots, n; j = 1, \ldots, q_i; k = 1, \ldots, r_i\} \tag{36}$$

where N'_{ijk} denotes the number of occurrence of the configuration $(V_i = v_{ik} | \Gamma_i^+ = \gamma_{ij}^+)$ for variable V_i given its parents Γ_i^+ in the whole data set of past collections, which we may denote by $\mathbf{D}'$. N'_{ijk} in the past data set $\mathbf{D}'$ has the same role as N_{ijk} in the present data set $\mathbf{D}$. With this form of prior information, it can be derived that

$$\begin{aligned} P(\mathbf{D}|S) \quad = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \quad & \frac{(N'_{ij} + r_i - 1)!}{(N'_{ij} + N_{ij} + r_i - 1)!} \\ & \prod_{k=1}^{r_i} \frac{(N'_{ijk} + N_{ijk})!}{N'_{ijk}!} \end{aligned} \tag{37}$$

B. Computing Parameter Expectation $E(\theta_{ijk}|\mathbf{D}, S)$ from Complete Data

The expectation of parameters θ_{ijk} given the complete data set $\mathbf{D}$ and a structure S is defined as

$$\begin{aligned} E[\theta_{ijk}|\mathbf{D}, S] \quad = \quad & \int_{\theta_{ij1}} \ldots \int_{\theta_{ijr_i}} \theta_{ijk} \\ & f(\theta_{ij1}, \ldots, \theta_{ijr_i}|\mathbf{D}, S) d\theta_{ij1}, \ldots, d\theta_{ijr_i} \end{aligned} \tag{38}$$

where the θ_{ijk} satisfy the two constraints (32)-(33). we have

$$E[\theta_{ijk}|\mathbf{D}, S] = \frac{N_{ijk} + 1}{N_{ij} + r_i} \tag{39}$$

Similar to the derivation of equation (37), if the prior information about the parameters is available in the form of (36), and the Dirichlet prior distribution is taken, then we can obtain

$$E\left[\theta_{ijk}|\mathbf{D}, S\right] = \frac{N_{ijk} + N'_{ijk} + 1}{N_{ij} + N'_{ij} + r_i} \tag{40}$$

V. EM Algorithms for Estimating Parameters and Data Likelihood with Incomplete Data

Let D_l denotes the l-th complete data case in the complete data set $\mathbf{D}$. Suppose now an incomplete and soft data set $\mathbf{D}'$ is given, its l-th incomplete data case is D'_l. In this situation, we need the EM algorithm to estimate each parameter θ_{ijk} given a structure S and the incomplete data set $\mathbf{D}'$.

The EM algorithm for ML estimation of the parameters from incomplete data set $\mathbf{D}'$ is: for the current $(t+1)$-th iteration,

1) The E-Step computes the expectation of the sufficient statistics N_{ijk}

$$\begin{aligned} N_{ijk}^{(t+1)} \quad &= \quad E\left[N_{ijk}|\mathbf{D}', \Theta^{(t)}\right] \\ &= \quad \sum_{l=1}^{N} p(V_i = v_{ik}, \Gamma_i^+ = \gamma_{ij}^+ | D'_l, \Theta^{(t)}, S) \end{aligned} \tag{41}$$

where D'_l is the l-th data case in the provided data set $\mathbf{D}'$. The probability in the above equation can be evaluated using any general-purpose Bayesian network inference algorithm. This point is explained below.

2) The M-Step computes an ML estimation of parameters θ_{ijk} using the expected sufficient statistics $N_{ijk}^{(t+1)}$ as if they were actual sufficient statistics from a complete data set $\mathbf{D}$ corresponding to the incomplete and soft data set $\mathbf{D}'$. The result is

$$\begin{aligned} \theta_{ijk}^{(t+1)} \quad &= \quad \arg\max_{\theta_{ijk}} Q(\Theta|\Theta^{(t)}) \\ &= \quad \arg\max_{\theta_{ijk}} E\left[P(\mathbf{D}|\Theta)|\mathbf{D}', \Theta^{(t)}, S\right] \\ &= \quad \frac{N_{ijk}^{(t+1)}}{\sum_{r=1}^{r_i} N_{ijr}^{(t+1)}} \end{aligned} \tag{42}$$

A. MAP Estimation of Parameters for Multinomial Distribution

The EM algorithm for MAP estimation of parameters is different from the ML estimation in the M-Step. Assume the prior probabilities $p'(V_i = v_{ik}, \Gamma_i^+ = \gamma_{ij}^+)$ exist, and there were N' samples in all the previous data sets used to obtain these prior probabilities, then we can define the prior counts as

$$N'_{ijk} = N' p'(V_i = v_{ik}, \Gamma_i^+ = \gamma_{ij}^+) \tag{43}$$

With this form of prior information in which the counts N'_{ijk} are no longer integer, but real in general, and using the general form of Dirichlet distribution as conjugate prior with multinomial distribution, we obtain the MAP estimation of θ_{ijk} as

$$\theta_{ijk}^{(t+1)} = \frac{N'_{ijk} + N_{ijk}^{(t+1)} + 1}{\left(\sum_{r=1}^{r_i}(N'_{ijr} + N_{ijr}^{(t+1)})\right) + r_i} \tag{44}$$

B. Estimating the Incomplete-Data Likelihood $P(\mathbf{D}'|S)$

Under the assumptions of unrestricted multinomial distributions, parameter independence, Dirichlet priors, and complete data, the complete-data likelihood $P(\mathbf{D}|S)$ of (37) can be rewritten in a more general form as

$$P(\mathbf{D}|S) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{45}$$

where $\Gamma(\cdot)$ denotes the Gamma function. It is difficult to compute the incomplete-data likelihood $P(\mathbf{D}'|S)$. It is also impractical to compute the expectation of the complete-data likelihood given the incomplete data $E\left[P(\mathbf{D}|S)|\mathbf{D}'\right]$ by using equation (45). This is because equation (45) is derived by using Bayesian average over all the possible parameter configuration, but the computation of the sufficient statistics N_{ijk}'s with incomplete data does require a parameter configuration θ_{ijk}'s because both N_{ijk}'s and θ_{ijk}'s have to be computed using the iterative EM algorithm, and this algorithm requires a general Bayesian network inference for its E-Step in each iteration. Even so, this costly EM algorithm only yields one set of parameters θ'_{ijk} and one set of sufficient statistics N'_{ijk}. The disadvantage of this approach for computing $E\left[P(\mathbf{D}|S)|\mathbf{D}'\right]$ by integrating over all the parameter configurations is that there is no closed-form solution and thus this approach is impractical and may even be infeasible.

An alternative approach to the Bayesian averaging with incomplete data is to actually consider maximizing $P(\mathbf{D}, \Theta|S)$ with regard to an optimal parameter configuration Θ rather than purely computing $P(\mathbf{D}|S)$. In other words, we can use $\max_\Theta E\left[P(\mathbf{D}, \Theta|\mathbf{D}', S)\right]$ to replace

$$E\left[P(\mathbf{D}|\mathbf{D}', S)\right] = E\left[\int_{\Theta_S} P(\mathbf{D}, \Theta|\mathbf{D}', S)d\Theta\right] \tag{46}$$

Let $\hat{\theta}_{ijk}$ and $\hat{N}_{ijk}$ denote the optimal incomplete-data estimation of θ_{ijk}'s and N_{ijk}. Then the incomplete-data likelihood can be written as

$$P(\mathbf{D}', \hat{\Theta}|S) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\hat{\theta}_{ijk}^{\hat{N}_{ijk}} \cdot \prod_{i=1}^{n}\prod_{j=1}^{q_i} f(\hat{\theta}_{ij}) \tag{47}$$

In general, we can assume the Dirichlet prior for $f(\hat{\theta}_{ij})$,

$$\begin{aligned} f(\hat{\theta}_{ij}) &= f(\hat{\theta}_{ij1}, \ldots, \hat{\theta}_{ijr_i}) \\ &= \frac{(\hat{N}_{ij} + r_i - 1)!}{\prod_{k=1}^{r_i}\hat{N}_{ijk}!}\prod_{k=1}^{r_i}\hat{\theta}_{ijk}^{\hat{N}_{ijk}} \end{aligned} \tag{48}$$

where

$$\hat{N}_{ij} = \sum_{k=1}^{r_i}\hat{N}_{ijk} \tag{49}$$

VI. Concluding Remarks

The theory presented in this paper provides a complete formalism for learning the structure and parameters of an underlying Bayesian network from a given data set. This formalism is well founded on the basis of MAP and MDL criteria, and every part of it is computable in principle. However, it should be pointed out that an efficient implementation of this theory is nontrivial.

Acknowledgements

This work was partially supported by a grant "Learning Bayesian networks for knowledge discovery and data mining" of the National Natural Science Foundation of China when Heping Pan was a Professor of Wuhan University, China.

References

[1] C. David, M. Christopher, and H. David, "Large-sample learning of bayesian networks is np-hard," in *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, (San Francisco, CA), pp. 124–133, Morgan Kaufmann Publishers, 2003.

[2] D. Heckerman, "A tutorial on learning with Bayesian networks (revised)," Tech. Rep. MSR-TR-95-06, Microsoft Research, Advanced Technology Division, Microsoft Corporation, 1 Microsoft Way, Redmond, WA 98052, USA, March 1996.

[3] W. Buntine, "A guide to the literature on learning probabilistic networks from data," *IEEE Trans. Knowledge and Data Engineering*, vol. 8, no. 2, 1996.

[4] P. J. Krause, "Learning probabilistic networks," in *http://www.auai.org/auai-tutes.html*, 1998.

[5] M. I. Jordan, *Learning in Graphical Models*. The Netherland: Kulwer Academic Publishers, 1998.

[6] H.-P. Pan, "Fuzzy Bayesian networks - a general formalism for representation, inference and learning with hybrid Bayesian networks," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 14, no. 7, pp. 941–962, 2000.

[7] G. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309–347, 1992.

[8] D. Geiger and D. Heckerman, "A characterization of the Dirichlet distribution with application to learning Bayesian networks," in *Proc. Eleventh Conference on Uncertainty in Artificial Intelligence(UAI'95)*, 1995.

[9] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. River Edge, NJ: World Scientific, 1989.

[10] J. Suzuki, "A construction of Bayesian networks from databases based on an MDL scheme," in *Proc. Ninth Conference on Uncertainty in Artificial Intelligence(UAI'93* (D. Heckerman and A. Mamdani, eds.), (San Francisco), Morgan Kaufmann, 1993.

[11] W. Lam and F. Bacchus, "Learning Bayesian belief networks: An approach based on the MDL principle," *Computational Intelligence*, vol. 10, pp. 269–293, 1994.

[12] R. Bouckaert, "Properties of Bayesian network learning algorithms," in *Proc. Tenth Conference on Uncertainty in Artificial Intelligence(UAI'94)* (R. L. de Mantaras and D.Poole, eds.), (San Francisco), Morgan Kaufmann, 1994.

[13] N. Friedman and D. Koller, "Being bayesian about network structure: A bayesian approach to structure discovery in bayesian networks," in *Proc. Uncertainty in Artificial Intelligence*, pp. 201–210, 2000.

[14] D. Madigan and J. York, "Bayesian graphical models for discrete data," *Int. Statistical Review*, vol. 63, pp. 215–232, 1995.

[15] P. Giudici and P. Green, "Decomposable graphical gaussian model determination," *Biometrika*, vol. 86, pp. 785–801, 1999.

[16] P. Giudici, P. Green, and C. Tarantola, "Efficient model determination for discrete graphical models," 2000.

[17] B. Taskar, V. Chatalbashev, and D. Koller, "Learning associative markov networks," in *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, (New York, NY, USA), p. 102, ACM Press, 2004.

[18] N. Friedman, "The Bayesian structural EM algorithm," in *Proc. Fourteenth conference on Uncertainty in Artificial Intelligence(UAI'98)* (G. Cooper and S. Moral, eds.), (San Francisco), Morgan Kaufmann, 1998.

[19] W. H. Hsu, H. Guo, B. B. Perry, and J. A. Stilson, "A permutation genetic algorithm for variable ordering in learning bayesian networks from data," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2002)*, (New York, NY), 2002.

[20] H. Guo, B. B. Perry, J. A. Stilson, and W. H. Hsu, "A genetic algorithm for tuning variable orderings in bayesian network structure learning," in *Eighteenth national conference on Artificial intelligence*, (Menlo Park, CA, USA), pp. 951–952, American Association for Artificial Intelligence, 2002.

[21] L. Yin, C.-H. Huang, and S. Rajasekaran3, "Parallel data mining of bayesian networks from gene expression data," in *Poster Book of the 8-th Annual Int'l Conference on Research in Computational Molecular Biology (RECOMB 2004)*, pp. 122–123, 2004.

[22] A. J. Novobilski, J. A. Kline, and F. M. Fesmire, "Using a genetic algorithm to identify predictive bayesian models in medical informatics," in *The International Conference on Information Technology (ITCC)*, 2004.

[23] H.-P. Pan, "Super Bayesian influence networks (SBIN) for capturing stochastic chaotic patterns in multivariate time series," in *6th International Conference on Optimization Techniques and Applications (ICOTA 6)*, (Ballarat, Australia), 2004.

[24] J. Rissanen, "Modelling by shortest data description," *Automatic*, vol. 14, pp. 465–471, 1978.

[25] H.-P. Pan and W. Förstner, "An MDL-principled evolutionary mechanism to automatic architecturing of pattern recognition neural network," in *IEEE Proc. 11th Int. Conference of Int. Association of Pattern Recognition(IAPR)*, 1992.